Explainable AI for Science

Shilpika[1], Carlo Graziani[1], Coleton Grossman[1], Mihai Anitescu[12], Bethany Lusch[1], Venkatram Vishwanath[1], Michael E. Papka[1]

Argonne National Laboratory, The University of Chicago

Introduction: Model Interpretability

"Explainable AI (XAI)" founded by DARPA

- **Transparency:** Shows the internal logic of the model or explains outputs in an interpretable way.
- **Interpretability:** Makes it easier for humans to understand the *meaning* of a model's predictions.
- **Justification:** Provides human-readable reasoning behind decisions.
- **Traceability:** Allows tracking the data and processing steps that led to a decision.

Local vs. global explanations

Model-agnostic vs. Model-specific

Explanation coverage = region in input space where the explanation applies

Previous Work

Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F. and Kerren, A. (2020), The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. Computer Graphics Forum, 39: 713-756. https://doi.org/10.1111/cgf.14034

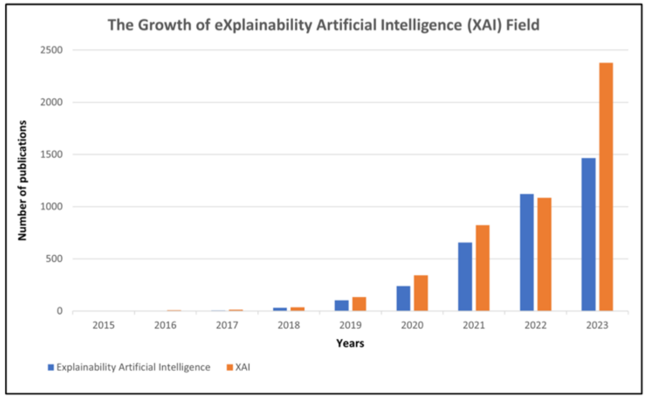


Figure 1: Research publications in the explainability of artificial intelligence (XAI) field during the last few years

Mersha, M., Lam, K., Wood, J., AlShami, A., & Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing, 599, 128111.

XAI method	Level		Abstraction		Dependency		
	Global	Local	Low	High	Data	Model	Domain
LIME [57]	○	●	○	●	●	○	○
ANCHORS [58]	○	●	○	●	●	○	○
CAV. [36]	○	●	●	●	●	●	●
e-LRP [6]	○	●	○	●	●	●	●
z-LRP [5]	○	●	○	●	●	●	●
DeepTaylor [50]	○	●	○	●	●	●	●
Saliency [68]	○	●	○	●	●	●	●
Gradient [3]	○	●	○	●	●	●	●
DeepLIFT [64]	○	●	○	●	●	●	●
grad*input [65]	○	●	○	●	●	●	●
Grad-CAM [61]	○	●	○	●	●	●	●
Occlusion [83]	○	●	○	●	●	●	●
SmoothGrad [70]	○	●	○	●	●	●	●
Integrated Grad [76]	○	●	○	●	●	●	●
DeConvNet [84]	○	●	○	●	●	●	●
Node-Link Vis [27]	●	●	●	○	●/○	●	○
Info Flow [80]	●	○	●	○	●/○	●	○
MinMax*	○	○	●	○	○	●	○
HistoTrend*	○	○	●	○	○	●	○
Dead Weight*	●	○	●	○	○	●	○
Saturated Weight*	●	○	●	○	○	●	○

Thilo Spinner, Udo Schlegel, Hanna Schäfer, & Mennatallah El-Assady (2019). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. CoRR, abs/1908.00087.

Toolbox	Publication	Code repository
InterpretML	Nori et al. (2019)	https://github.com/interpretml/interpret
iNNvestigate	Alber et al. (2019)	https://github.com/albermax/innvestigate
AI Fairness 360	Arya et al. (2019)	https://github.com/Trusted-AI/AIF360
explAIner	Spinner et al. (2020)	https://github.com/dbvis-ukon/explainer
FAT Forensics	Sokol et al. (2020)	https://github.com/fat-forensics/fat-forensics
Alibi	Klaise et al. (2021)	https://github.com/SeldonIO/alibi

Model-Agnosticism And Statistical Learning

Key Idea: Adopt a view of AI that abstracts away the model

Focus not on what the model *is*, but rather on what it *does*.

All statistical learning features a *training stage* and an *inference stage*

Training:

- Take a set of data, and infer an approximation to the statistical distribution from which the data was sampled;
  - “Data" could be images, weather states, protein structures, text...
- Simultaneously, optimize some decision-choosing rule in a space of such decisions, exploiting the learned distribution;
  - “Decision" could be the forecast of a temperature, or a label assignment to a picture, or the next move of a robot in a biochemistry lab, or a policy, or a response to a text prompt...

Inference:

When presented with a new set of data samples,

- produce the decisions appropriate to each one, using facts about the data distribution and about the decision optimality criteria inferred in training.

This view of AI encourages a model-agnostic outlook, which emphasizes data distributional structure and decision-space optimality, and diverts attention away from model computational structure.

Probabilistic Attribution Score: AS

$$A_\mu \equiv \log \frac{\Pr(\boldsymbol{R}|\boldsymbol{P})}{\Pr(\boldsymbol{R}|\boldsymbol{P} - P_\mu)},$$

The basic idea is that if we have a prompt **P** that induces a response **R** sampled from the distribution **P (R|P)** (where **P** and **R** are token sequences), then we would like a measure of attribution of the importance to the **μ**-th prompt token **Pμ** to **R.**

Probabilistic Attribution Score: AS

$$A_\mu \equiv \log \frac{\Pr\left(\boldsymbol{R}|\boldsymbol{P}\right)}{\Pr\left(\boldsymbol{R}|\boldsymbol{P}-P_\mu\right)},$$

$$\begin{aligned}\Pr\left([L_\mu,\ldots,L_{\mu+\nu}]\,|\,[L_1,\ldots,L_{\mu-1}]\right) &= \Pr\left(L_\mu|\,[L_1,\ldots,L_{\mu-1}]\right)\\ &\quad\times\Pr\left(L_{\mu+1}|\,[L_1,\ldots,L_\mu]\right)\times\ldots\\ &\quad\ldots\times\Pr\left(L_{\mu+\nu}|\,[L_1,\ldots,L_{\mu+\nu-1}]\right).\end{aligned}$$

The basic idea is that if we have a prompt **P** that induces a response **R** sampled from the distribution **P (R|P)** (where **P** and **R** are token sequences), then we would like a measure of attribution of the importance to the **μ**-th prompt token **Pμ** to **R.**

Probabilistic Attribution Score: AS

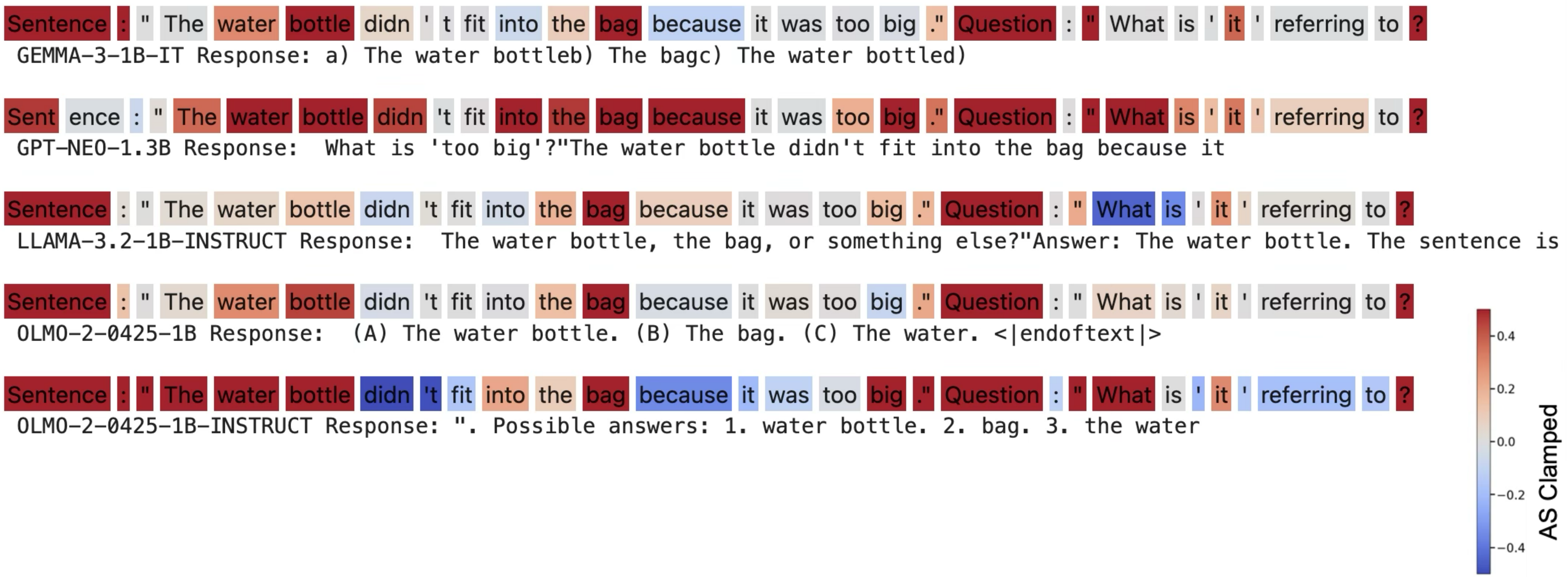
$$A_\mu \equiv \log \frac{\Pr\left(\boldsymbol{R}|\boldsymbol{P}\right)}{\Pr\left(\boldsymbol{R}|\boldsymbol{P}-P_\mu\right)},$$

$$\begin{aligned} D &\equiv \Pr\left(\{\boldsymbol{R}=\boldsymbol{r}\}\,|\,\{\boldsymbol{P}-P_\mu=\boldsymbol{p}-p_\mu\}\right) \\ &= \Pr\left(\{\boldsymbol{R}=\boldsymbol{r}\}\,|\,\left\{\boldsymbol{P}^{<\mu}=\boldsymbol{p}^{<\mu}\right\}\cap\left\{\boldsymbol{P}^{>\mu}=\boldsymbol{p}^{>\mu}\right\}\right) \end{aligned}$$

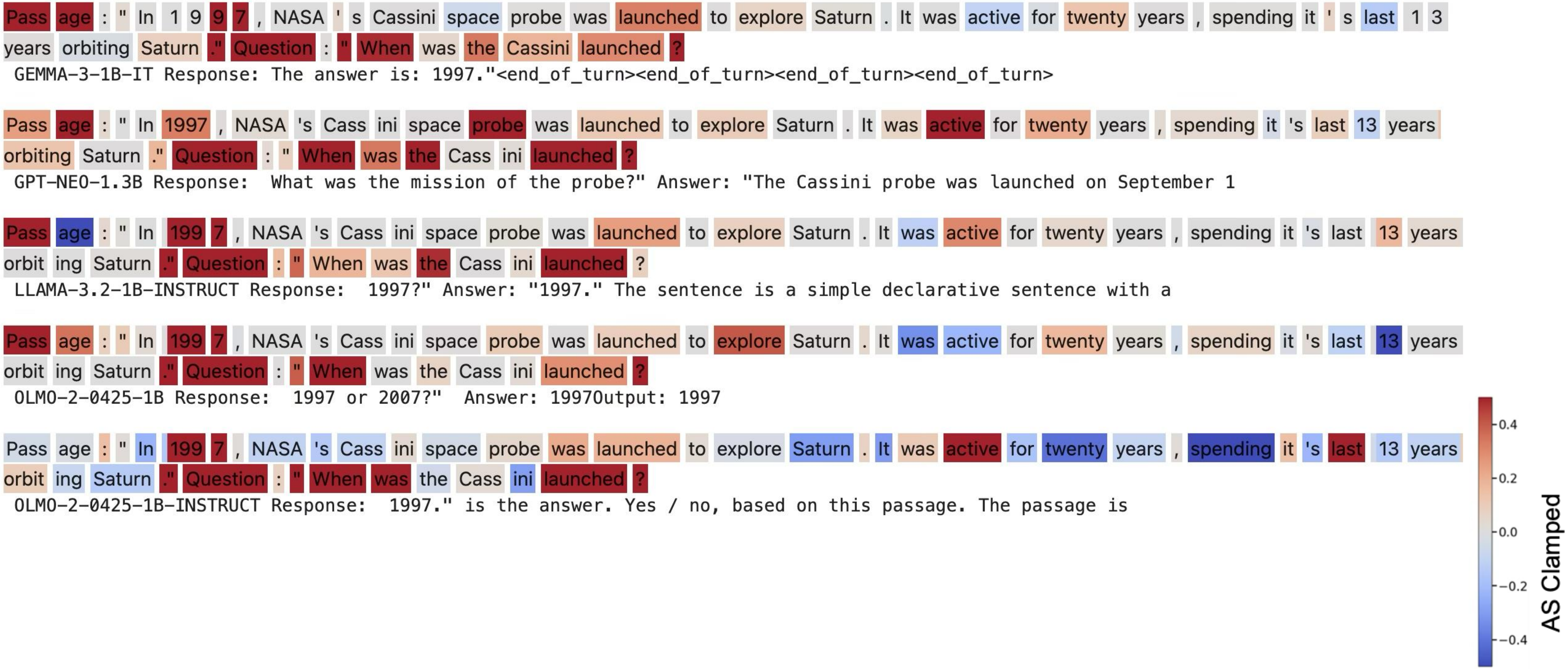
The basic idea is that if we have a prompt **P** that induces a response **R** sampled from the distribution **P (R|P)** (where **P** and **R** are token sequences), then we would like a measure of attribution of the importance to the **μ**-th prompt token **Pμ** to **R.**

D is the denominator of Aμ.

Probabilistic Attribution Score: AS

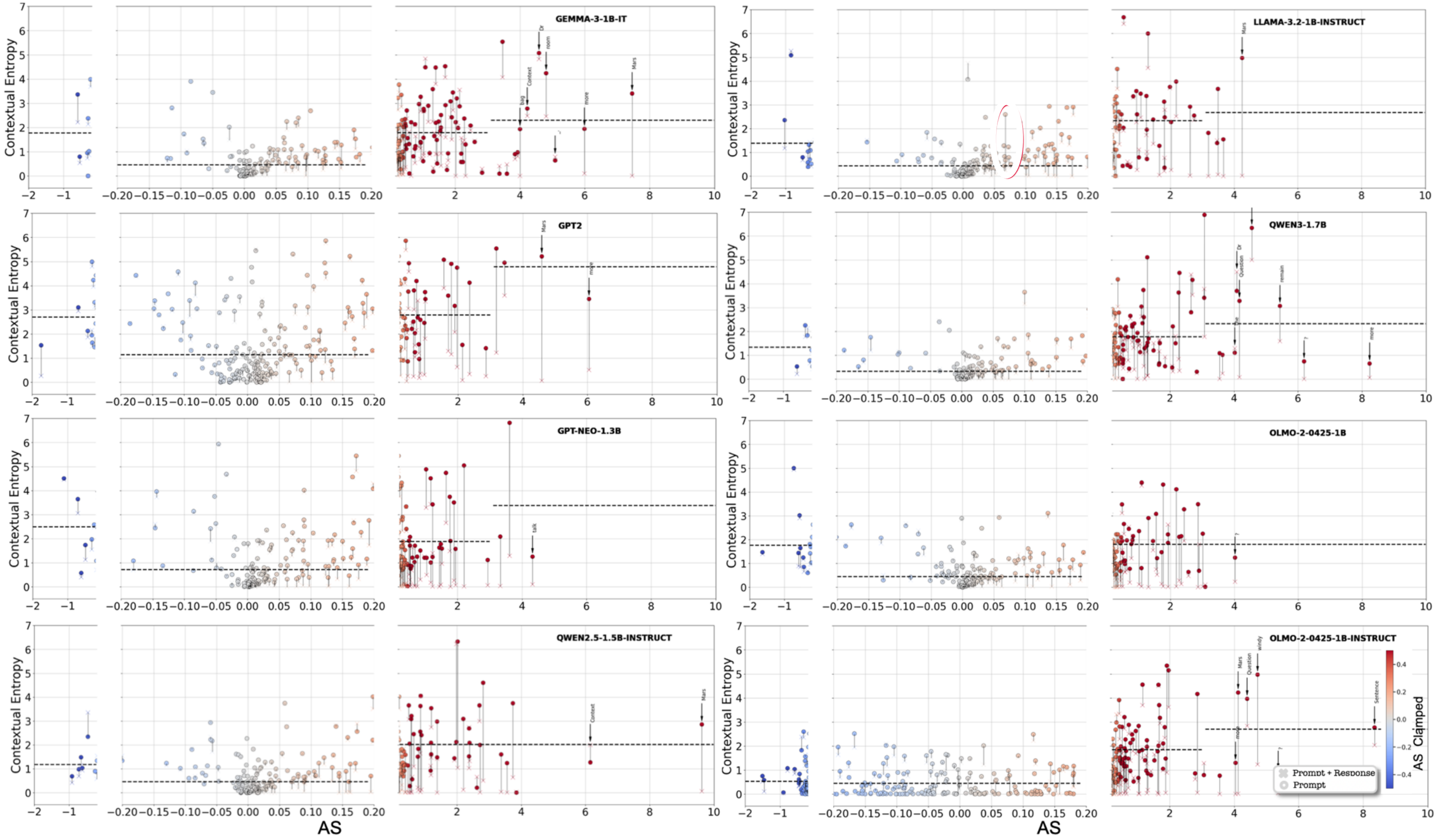


Probabilistic Attribution Score: AS



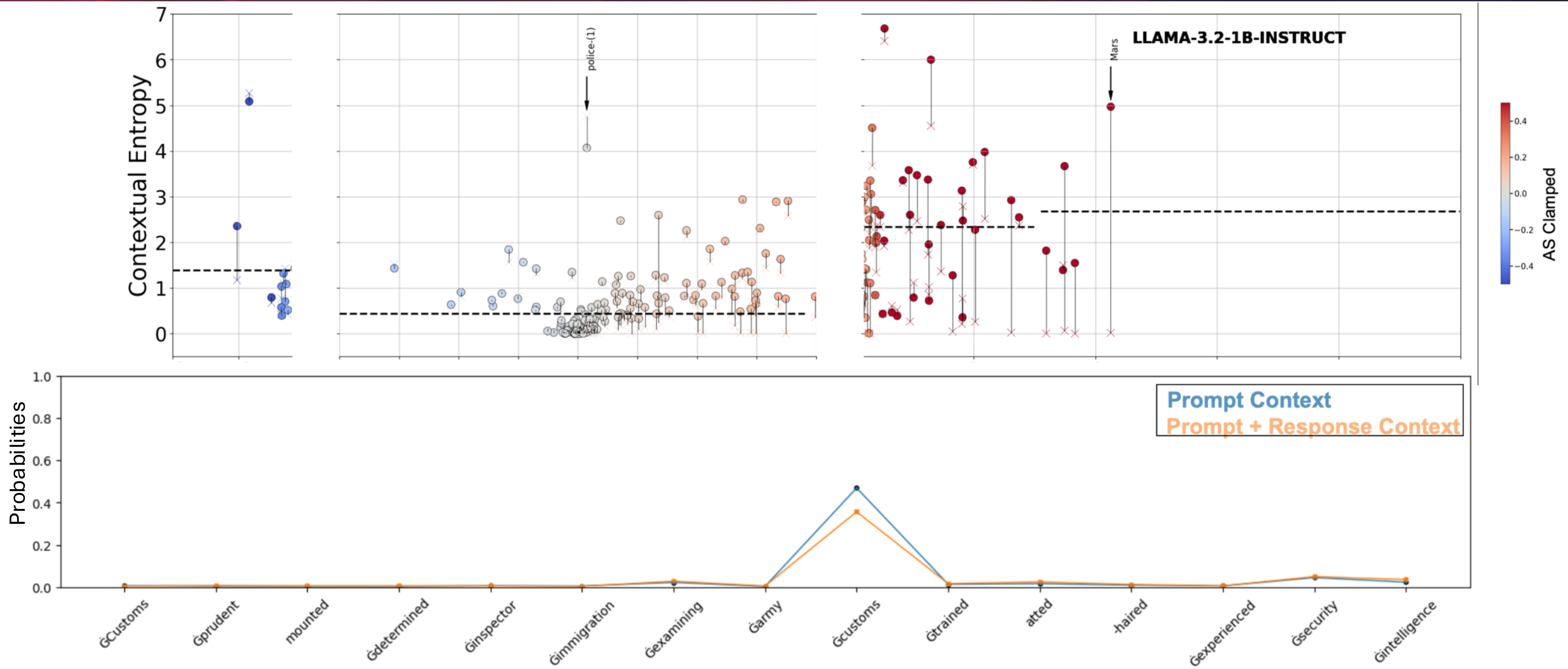
Contextual Entropy vs AS - Greedy

- Contextual Entropy: Entropy of the distribution of each token conditioned on the rest of the context: with or without the response context.
- Low contextual entropy means that given the context (◦ or ×) the model prefers fewer replacement tokens for the given token.
- High contextual entropy corresponds to more uncertainty in potential token replacements candidates.



Contextual Entropy with Prompt and Prompt + Response Context

- Token: Police
- Contextual entropy is higher at near-zero AS.
- KL Divergence is low = 0.0075.
- Greedy sampling

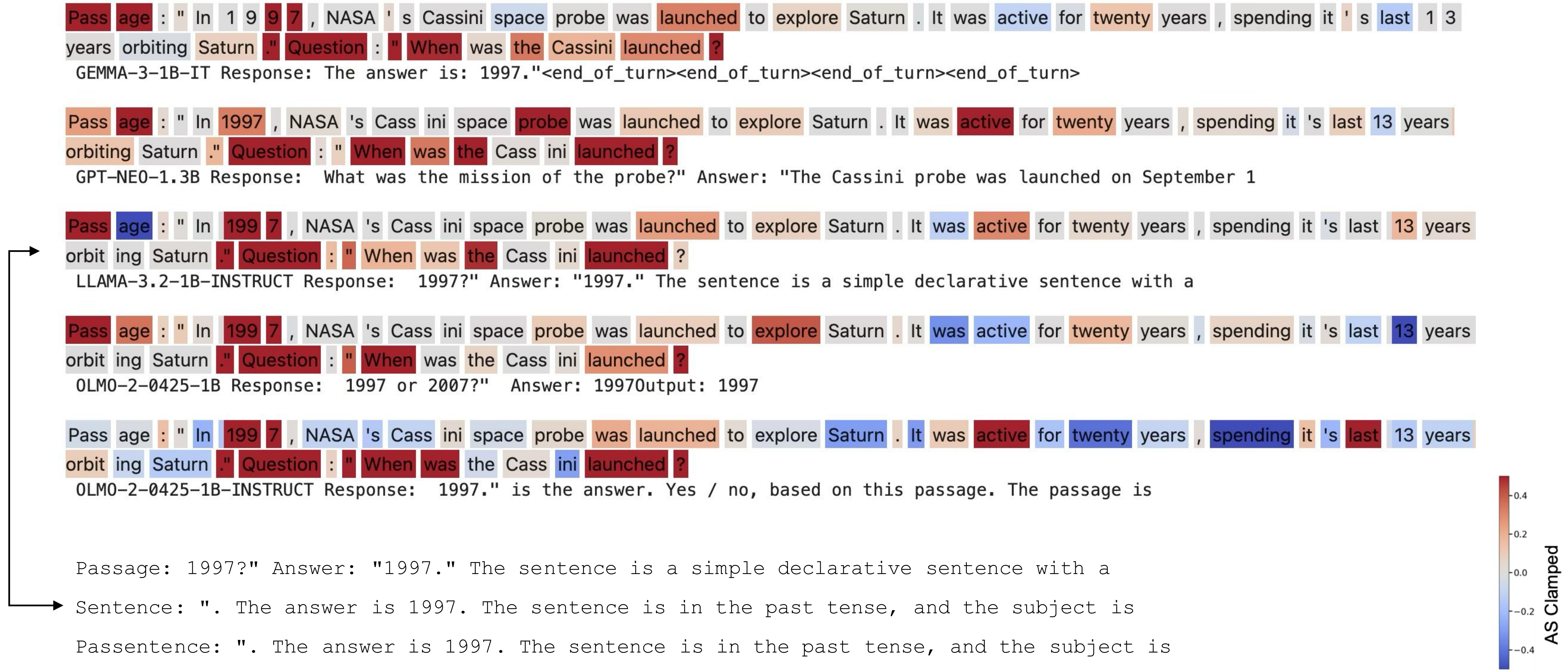


Prompt: The ***police*** officer walked into the room, carefully examining every clue. The information was sparse, but he believed the clues would lead him to the truth. He looked at the doorknob, noticing something odd about it. He made a note to check it more

Response: closely.

He walked over to the door and examined the doorknob more closely. It was a…

AS: Prompt Engineering



AS: Prompt Engineering

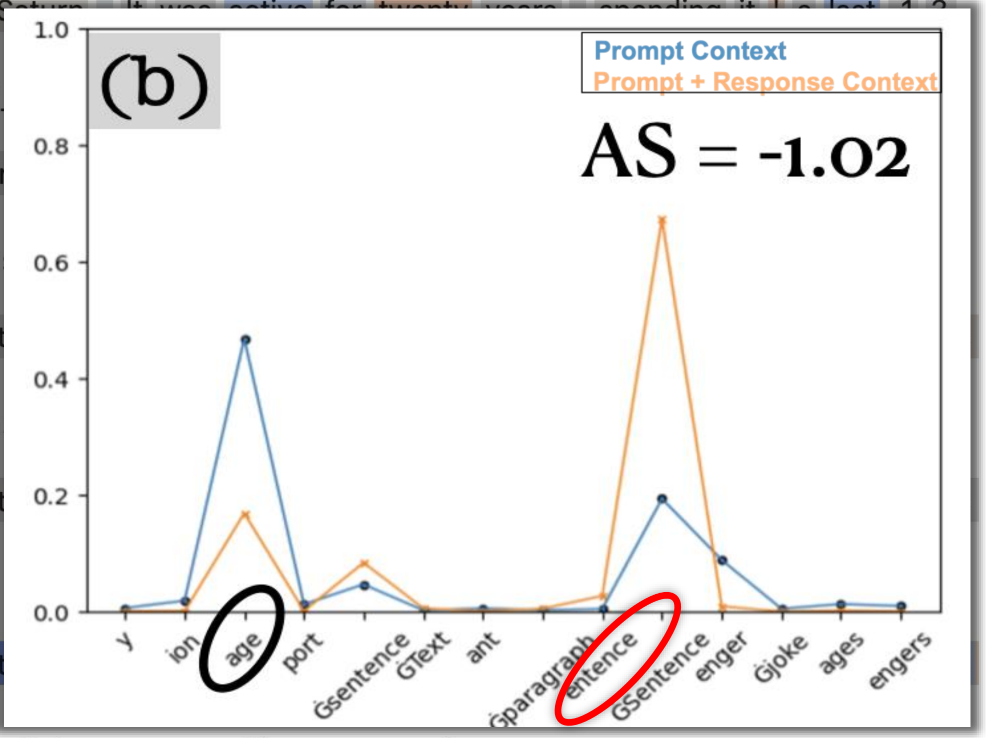
Pass age : " In 1 9 9 7 , NASA ' s Cassini space probe was launched to explore
years orbiting Saturn ." Question : " When was the Cassini launched ?
GEMMA-3-1B-IT Response: The answer is: 1997."<end_of_turn><end_of_turn><end

Pass age : " In 1997 , NASA 's Cass ini space probe was launched to explore Satu
orbiting Saturn ." Question : " When was the Cass ini launched ?
GPT-NEO-1.3B Response: What was the mission of the probe?" Answer: "The Ca

Pass age : " In 199 7 , NASA 's Cass ini space probe was launched to explore Sat
orbit ing Saturn ." Question : " When was the Cass ini launched ?
LLAMA-3.2-1B-INSTRUCT Response: 1997?" Answer: "1997." The sentence is a s

Pass age : " In 199 7 , NASA 's Cass ini space probe was launched to explore Sat
orbit ing Saturn ." Question : " When was the Cass ini launched ?
OLMO-2-0425-1B Response: 1997 or 2007?" Answer: 1997Output: 1997

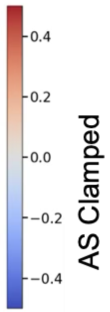
Pass age : " In 199 7 , NASA 's Cass ini space probe was launched to explore Sat
orbit ing Saturn ." Question : " When was the Cass ini launched ?
OLMO-2-0425-1B-INSTRUCT Response: 1997." is the answer. Yes / no, based on this passage. The passage is



Passage: 1997?" Answer: "1997." The sentence is a simple declarative sentence with a

Sentence: ". The answer is 1997. The sentence is in the past tense, and the subject is

Passentence: ". The answer is 1997. The sentence is in the past tense, and the subject is

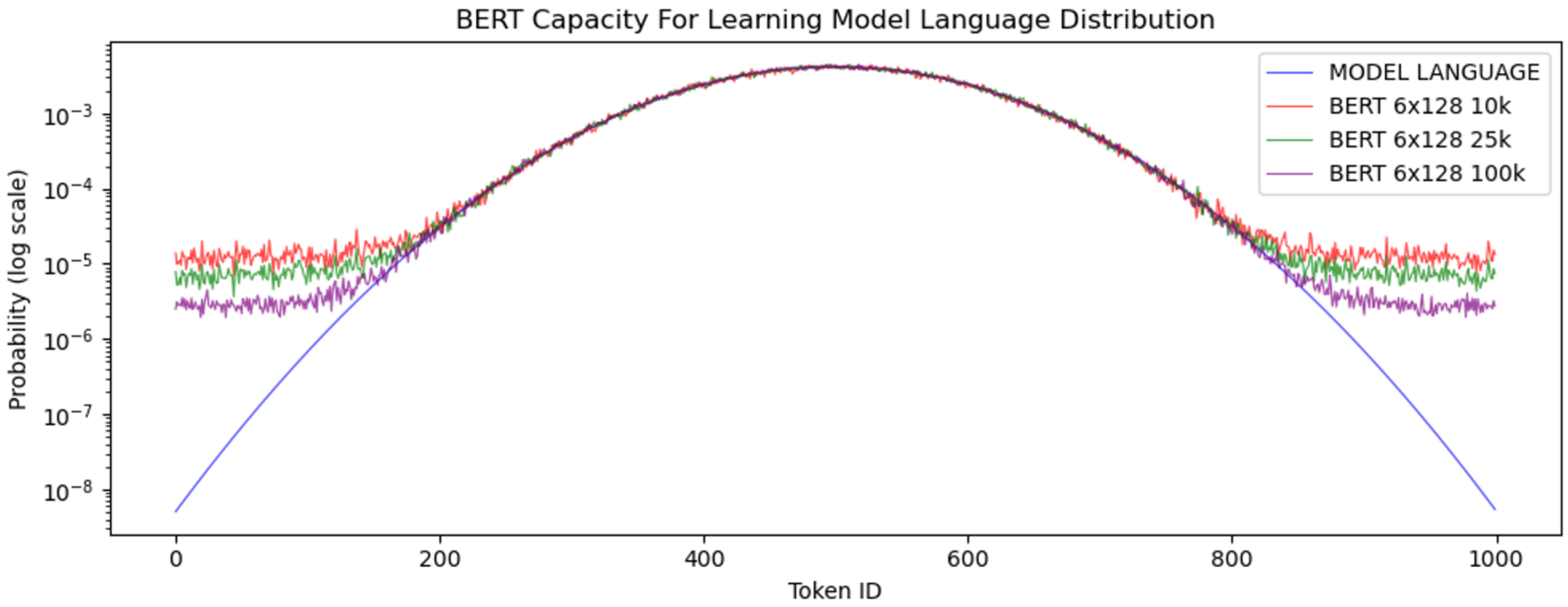


Model Language as a Testbed for Large Language Models

- Gaussian mixture of dialects such that we sample a dialect with weights, *w,* and then sample from the corresponding dialect matrix (also multivariate) and adding together the vectors.
- We set it so that 10% of the tokens differ between the common language and selected dialect.
- We tokenize along the real number line with bounds set to the bounds of our Gaussian mixture of dialects.
- Thus, we can calculate the next token probabilities of our model language.

BERT Capacity for Learning Model Language

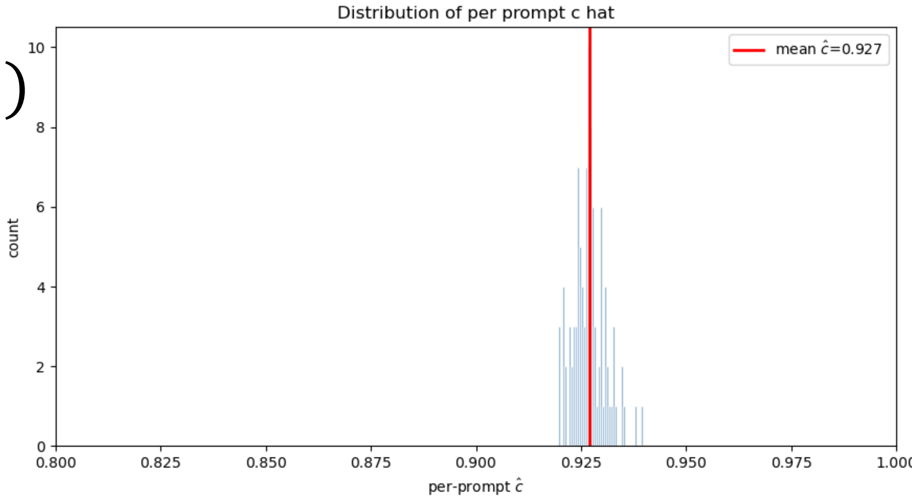
- We use a mini-BERT with 6 hidden layers and 128 hidden size.
- We plot BERT next-token probabilities compared to model next token probabilities and show it learns the distribution further as training vectors *n* increase.



Hypothesis for Weight Shifting

A model fine-tuned on a single dialect must be a convex linear combination of a pre-trained model and a model fine-tuned only on that single dialect.

- $cP(t_n|T_{<n}, M_P) + (1-c)P(t_n|T_{<N}, M_s) = P(t_n|T, M_F)$
- Rearranging to regress for c using OLS we have,
- So, $\hat{c} = \frac{\sum_n (P_P - P_S)(P_F - P_S)}{\sum_n (P_P - P_S)^2}$
- P = pretrained, F = finetuned, S = trained on a single dialect

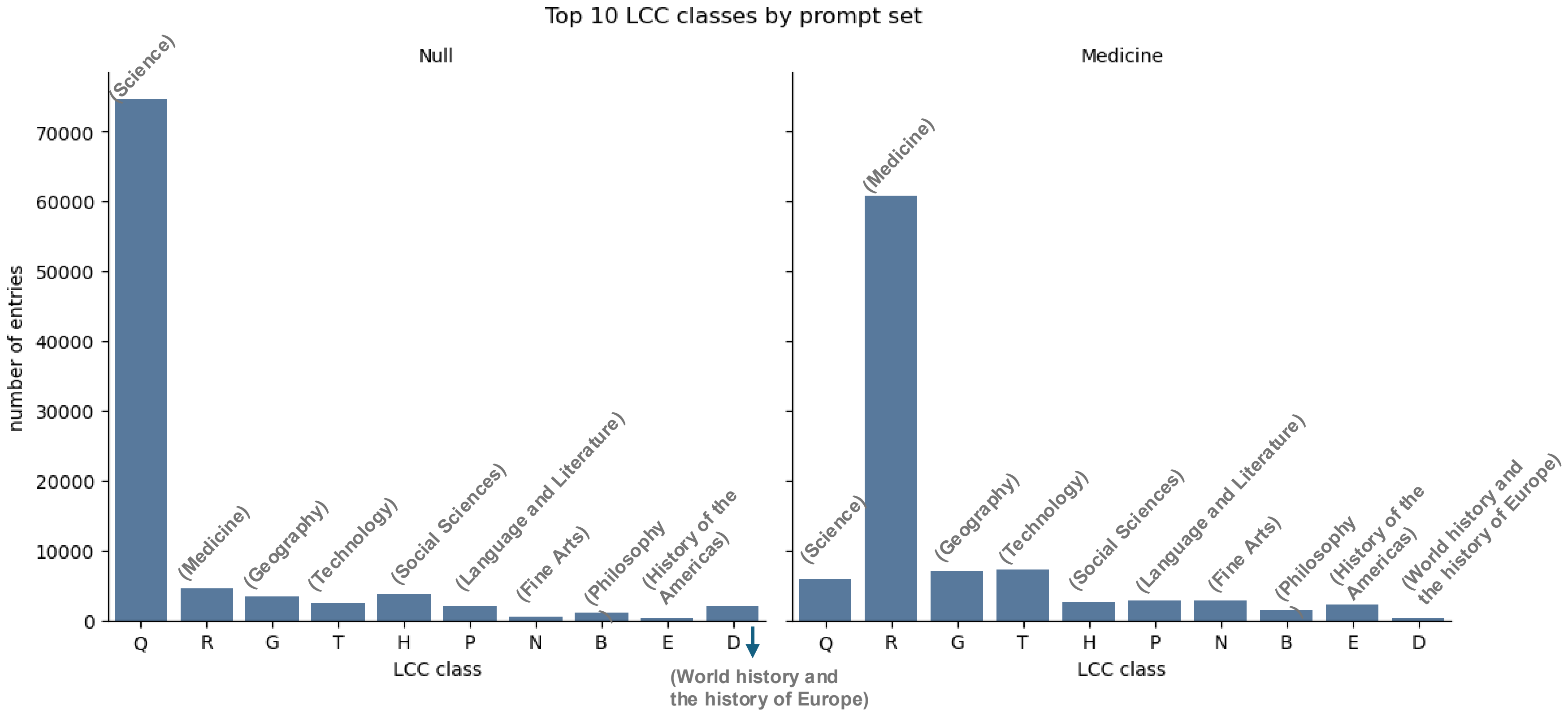


Null Prompting

- Null prompted Llama 3-8B model 100,000 times
- Prompted Llama 3-8B model with prompt “medicine” 100,000 times
- Response length set to128 tokens
- Had Claude classify the responses according the the Library of Congress classification system :
  - class(First Letter – “Q”= Science)
  - subclass(Second letter – “QA” = Mathematics)
  - subcategory(number range – QA76.75 = Computer Software)

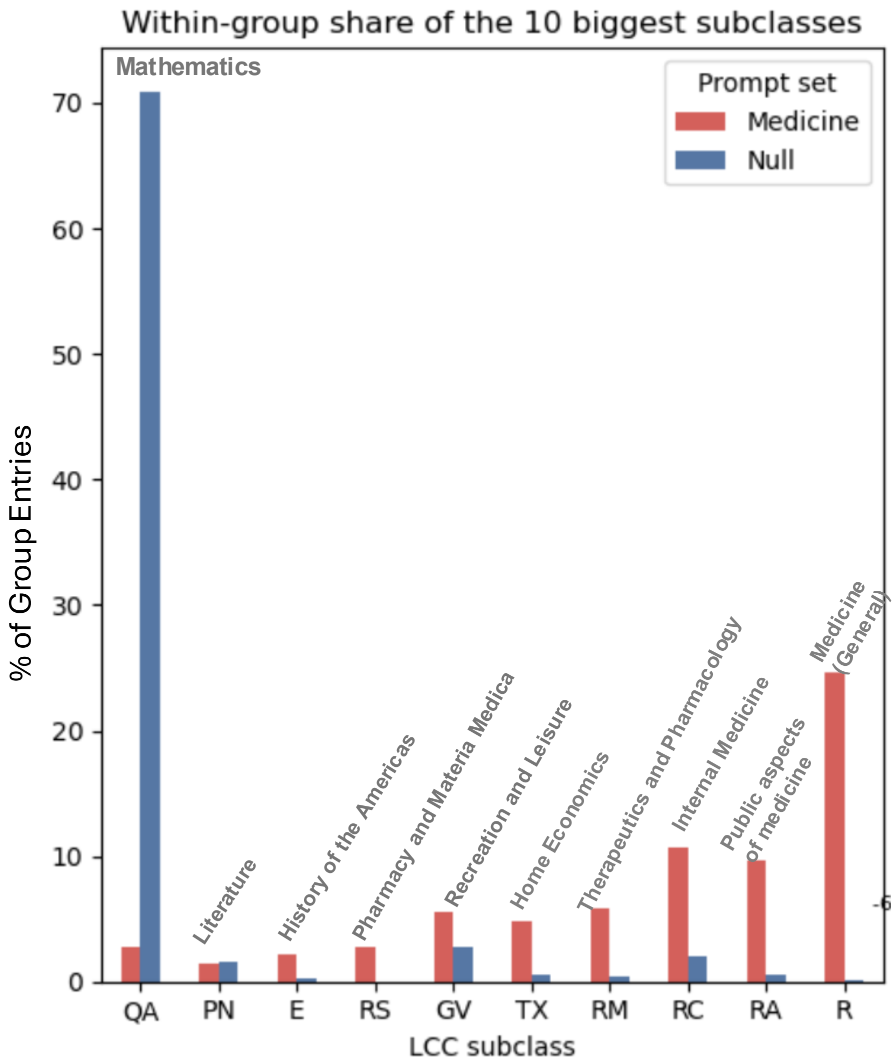
Shifting of Weight

We essentially have a picture of what distribution that Claude is sampling from in its responses based on its classifications and see a shift in weights.



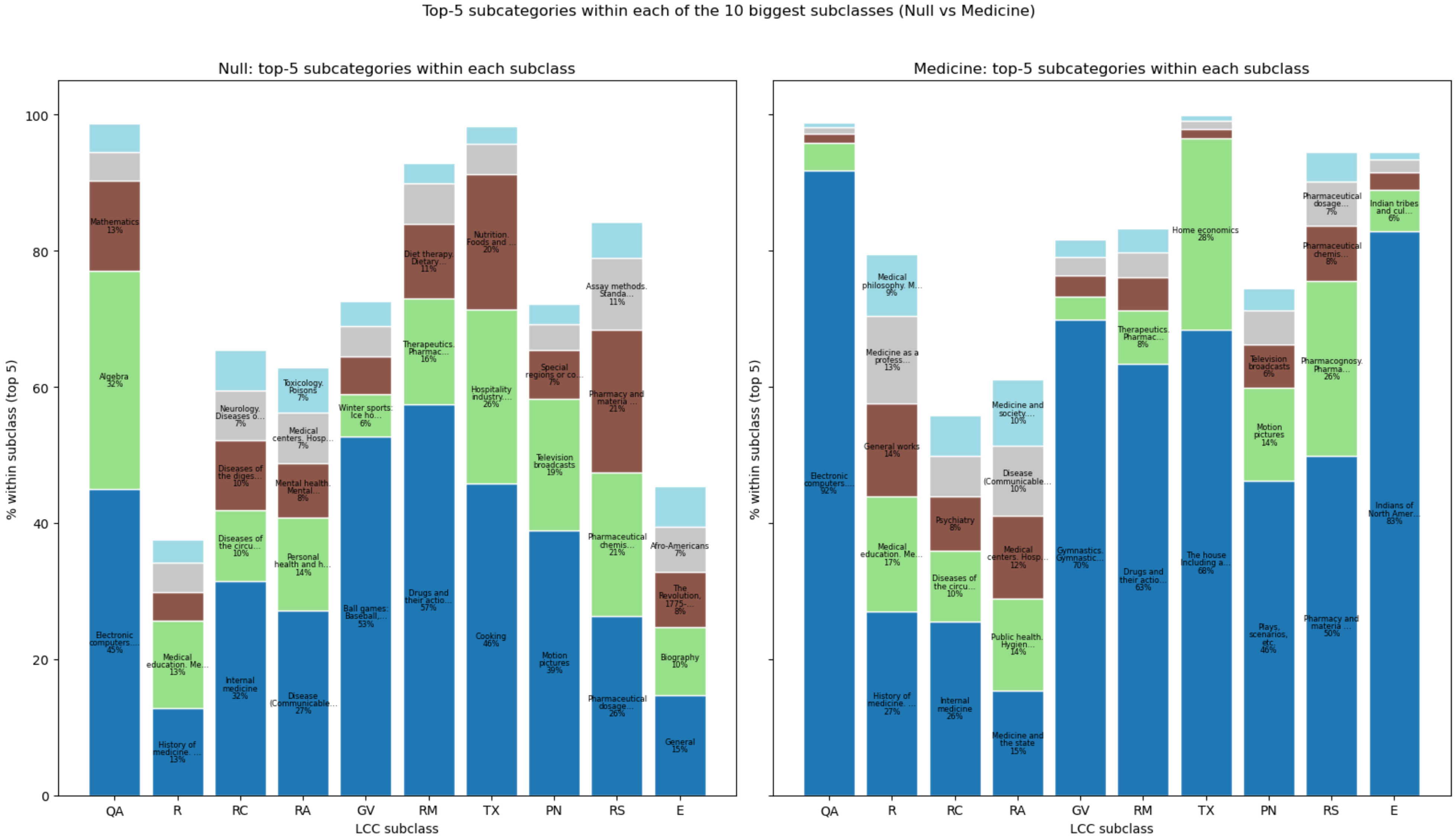
Shift within Subclass

- Shifts within each subclass
- Not just shift in class sampling
- This graph shows how much of these big subclasses made up the overall class distributions and the shift between them

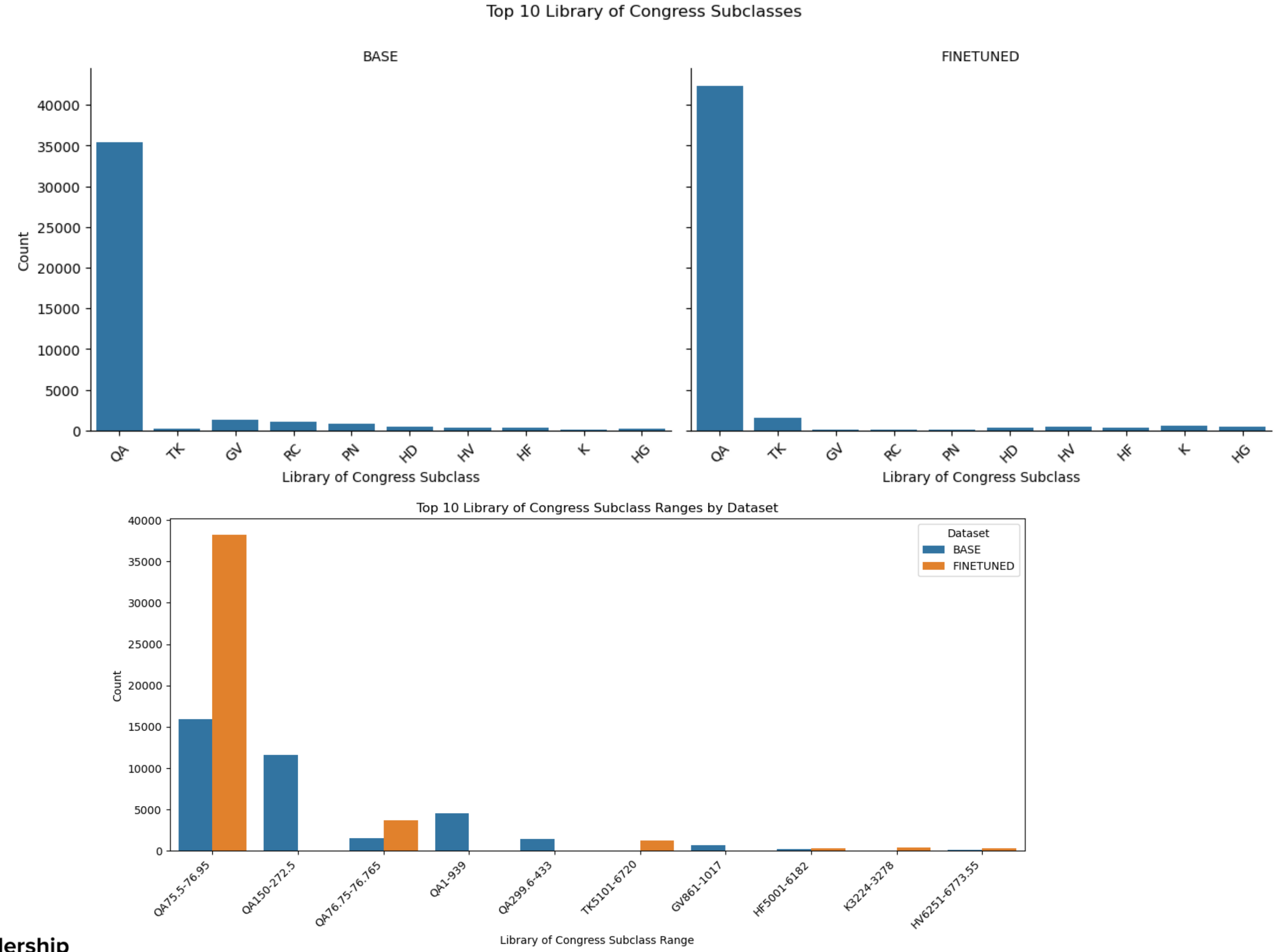


Shift within Subclass

We also see significant subcategory shift within each subclass between the prompts

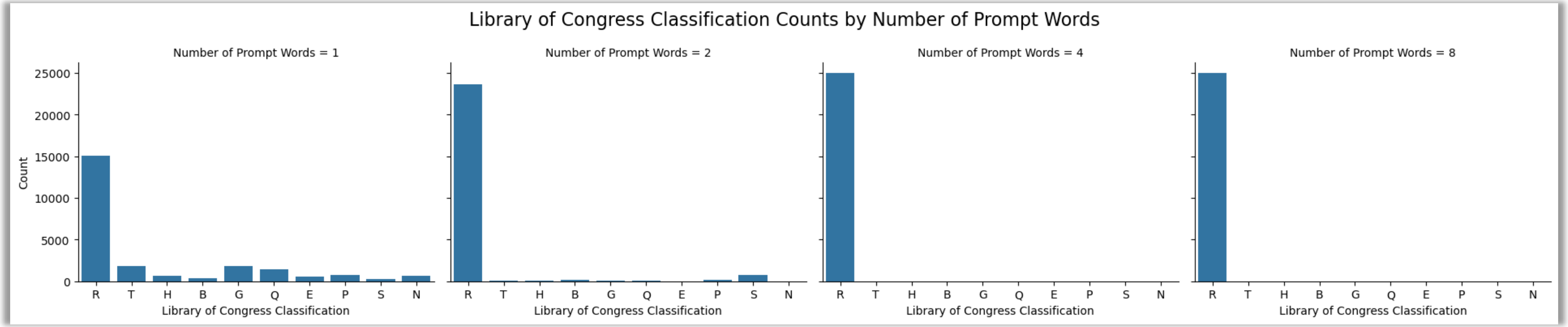


Shifting of Weight – Base and Fine-Tuned



Shifting of Weight – Prompt Length

We essentially have a picture of what distribution that Claude is sampling from in its responses based on its classifications and see a shift in weights



R: Medicine
T: Technology
H: Social Sciences
N: Fine Arts

B: Philosophy
G: Geography
Q: Science
E: History Of The Americas

P: Language And Literature
S: Agriculture

Conclusion

- Model-agnostic explanations
- Test how well an LLM converges to the underlying distribution of the training data
- Non-parametric linking first principles of probability theory to token attribution
- Can be used to test other XAI method, since sensitivity is inherently build-in the attribution method
- Compare different LLM models
- Reinforcement learning and its effects on the entropy of the model responses and catastrophic forgetting

References & Reading

1. Shilpika Shilpika, Carlo Graziani, Bethany Lusch, Venkatram Vishwanath, & Michael E. Papka. (2026). Probabilistic Attribution For Large Language Models.
2. Mersha, M., Lam, K., Wood, J., AlShami, A., & Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing, 599, 128111.
3. Shankar, R., Goh, Z., Devi, F., & Xu, Q. (2025). A systematic review of explainable artificial intelligence methods for speech-based cognitive decline detection. NPJ Digit. Med., 8(1), 724.
4. D. Bhati, M. Amiruzzaman, Y. Zhao, A. Guercio and T. Le, "A Survey of Post-Hoc XAI Methods From a Visualization Perspective: Challenges and Opportunities," in IEEE Access, vol. 13, pp. 120785-120806, 2025, doi: 10.1109/ACCESS.2025.3581136.
5. Yang, W., Wei, Y., Wei, H. et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Hum-Cent Intell Syst 3, 161–188 (2023). https://doi.org/10.1007/s44230-023-00038-y
6. Thilo Spinner, Udo Schlegel, Hanna Schäfer, & Mennatallah El-Assady (2019)., "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 1064-1074, Jan. 2020, doi: 10.1109/TVCG.2019.2934629.
7. Billingsley, P., Probability and Measure. Wiley Series in Probability and Statistics. Wiley 1995 ISBN: 9780471007104
8. Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F. and Kerren, A. (2020), The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. Computer Graphics Forum, 39: 713-756. https://doi.org/10.1111/cgf.14034

ARGONNE
ATPESC2026
EXTREME-SCALE COMPUTING

ARGONNE TRAINING PROGRAM ON EXTREME-SCALE COMPUTING

Produced by Argonne National Laboratory, a U.S. Department of Energy Laboratory managed by UChicagoArgonne, LLC under contract DE-AC02-06CH11357.

Special thanks to the National Energy Research Scientific Computing Center (NERSC) and Oak Ridge Leadership Computing Facility (OLCF) for the use of their resources during the training event.